Question 1.3: What is the granularity of our dataset? Think of what each row represent. Choose 3 arbitrary columns you find interesting and explain how they help you understand the dataset's granularity. One of them should identify the *primary key* of this dataset. (Note that the primary key can be a combination of 2 or more columns.)

Hint: You can use `pandas.Series.value_counts` and/or `pandas.Series.unique`.

Each row represents one recorded pice of data for a specific husehold, given by the household id: `a1_hh_id`. The `a2_spring_id` is the water spring that the household obtains its water from.

Question 2.1: What are the main parts of the survey? In this question, list out each section denoted by a letter an explain in 1 sentence what you believe to be its significance. We'll start you off with two:

- Section A: Introduction with general respondent and interview round information and consent.
- Section B: Characteristics of respondent. Filled out once during the survey rounds (if respondent stays the same).

- Section C: Current health of the child/children.
- Section D: History of the child, specifically medical issues.
- Section E: Test to see what ailments are already affecting the child.
- Section F: Gaining more isnight on teh child's history.
- Section G: Asking how and where they store their water. In containers with other chemicals or liquids?
- Section H: Taking water samples. This is one of the more crucial parts of the survey as it is measuring water qaulity.
- Section I: Rewarding the participants, specifically with medication that can help those who becamse sick from water source.
- Section J: Ensures that the survey was filled out to the max and little to no blanks.
- Section K: More specification on how well the survey was conducted and how much data was gathered.

Question 2.2: After your first glance of the survey, what do you deem to be the most important “datapoints” collected that are relevant to the paper’s research hypothesis? You can either refer to specific questions and columns.

Hint: this paper focuses on the prevalence of diarrhea across treatment and control groups.

I think the most important data points collected are the current health of the child as well as the actual samples of the springs.

Question 2.3: Outside of the paper's "sphere of research interest", what would be interesting datapoints to analyse further? This is an open-ended question, and we suggest you form a short research question and how you would use the data from the survey.

Something that would be of further interest would be the parent's health. This could show generational issues with the springs or if something recent popped up that made it contaminated.

Question 3.3: In the text cell below, share an observation from the plot and what you believe potential causes of the variation of participating households in each round could be.

The earlier rounds saw much more data. This may be because of survey fatigue.

Question 3.6: Do you observe any particular trends in the reported past 7-day prevalence of child diarrhea across the survey rounds? Think of how its prevalence changes relative to previous survey rounds. Furthermore, take note of potential reasons for the trends you are observing.

The households with higher id's are shwoing to be more negative, which is a good thing.

Question 3.9: Choose one of the plots above and thoroughly reflect on a set of observations in a few sentences. Can you think of why disease prevalence is steadily declining as the number of survey rounds increase? And, what could have caused the sudden uptick in the last rounds? (Hint: Revisit the lecture slides).

Cough and chest noise seem to be correlated as we see a rise in both at around the same time and same rounds. This is probably because whatever disease is causing each symptom could be the same disease or that cough and chest noise often come hand in hand.

Question 4.2: Look at the graph above. The red points are the corresponding control groups 99 and 161. How different are these from the normal group quantitatively? (Feel free to just eyeball it or write some code) Are you surprised by your findings?

These points are pretty different from the rest. We can see how much further they are from the line that is plotted, showing the disparity.

Question 5.2: What does each row of `hh_wg` contain? What does it say about the granularity (or the level of aggregation)? How does it compare to the dataframe used in phase 1-4?

Each row contains data on a specific quiz taken. This is similar to the first df, which had specific household id's of children who took part in the survey.

Question 5.6: Which of the two Wateguard columns inform us whether or not a given household is in the *treatment* or *control* group? Which column stores our *outcome* variable?

The `validated_wg` and `promoted_wg` columns show treatment and control groups.

Question 5.10 (Extra Credit): Interpret your findings using the Sign, Significance, and Size framework.

Hint: If you're new to interpreting `statsmodels` summaries, you might find this blog post helpul.

Type your answer here, replacing this text.